A. Lavrentiev

SYNTACTIC REFERENCE CORPUS OF MEDIEVAL FRENCH (SRCMF)[1]

1. Project overview

The original goal of the SRCMF project[2] was to produce a syntactically annotated reference corpus for medieval French (9th – 13th centuries) by adding syntactic markup to the texts of two major medieval French corpora, the Base de Français Médiéval[3] and the Nouveau Corpus d'Amsterdam[4]. Such a corpus would be unique for Medieval French, although some Old French texts have been annotated within the *Modéliser le changement: les voies du français* project at the University of Ottawa[5].

The elaboration of a consensual linguistic model dealing with various constructions occurring in medieval texts, the development of software for computer-assisted annotation and verification and the annotation itself have turned out to be more time consuming than initially estimated. Therefore, only a dozen texts (approx 300 ths tokens) will probably be fully processed by the end of the project.

However, the project has already produced at least three kinds of useful output. First of all, discussion relating to the linguistic model and of the model testing on actual text data have provided interesting insights into some problems of the history of the French language[6].

---

[1] The project is funded by the Agence Nationale de la Recherche, France and the Deutsche Forschungsgemeinschaft, Germany (2009–2011).

[2] **https://listes.cru.fr/wiki/srcmf**

[3] **http://bfm.ens-lyon.fr**

[4] **http://www.uni-stuttgart.de/lingrom/stein/corpus**

[5] **http://www.voies.uottawa.ca**

[6] *Lavrentiev A.* La «phrase» en français médiéval: une réalité ou une reconstruction artificielle? // Actes du 2e Congrès mondial de linguistique

Secondly, NotaBene – open-source RDF annotation software developed by the project – has become a valuable tool reusable for other purposes [1]. Finally, the size of annotated corpus should be sufficient to serve as a «gold standard» for developing and training automated syntactic parsers for medieval French.

The project is carried out by a joint French (Paris and Lyon) and German (Stuttgart) team of university faculty members, researchers and engineers. Four post-doctoral fellows have been fully or partially funded by the project. The work is coordinated by Dr S. Prévost (Lattice research lab, Paris) and by Prof Dr A. Stein (Stuttgart University).

2. Annotation workflow

Each text is independently annotated by two experts using NotaBene (fig. 1). Later, the same software is used to compare their output. After the elimination of obvious errors, the remaining differences (due to comprehension difficulties, text ambiguity or differing interpretations of the linguistic model) are either submitted for general discussion on the forum of the project or directly passed

---

française. New Orleans, 12–15 July 2010. P. 277–289. **http://www.linguistiquefrancaise.org**; *Mazziotta N.* Traitement de la coordination dans le Syntactic Reference Corpus of Medieval French (SRCMF) // Actes du XXVIe CILPR. Valencia, 6–11 Septembre 2010. Forthcoming; *Glikman J., Mazziotta N.* Représentation de l'oral et syntaxe dans la prose du Queste del saint Graal (1225-1230) // International conference «Représentation du sens en linguistique V». Chambéry, 25–27 May 2011. Oral presentation.

[1] *Mazziotta N.* Building the Syntactic Reference Corpus of Medieval French using NotaBene RDF Annotation Tool // Proceedings of the Fourth Linguistic Annotation Workshop. Stroudsburg, PA, USA. P. 142–146. The software is freely available at **http://sourceforge.net/projects/notabene**. Please note that the software is still in alpha version, and the readers are invited to contact the author if intending to use it.

over to the project supervisors who make the final decision. In the early stages of the project these discussions allowed the clarification of some aspects of the linguistic model. At present, four annotators are working on the project on a full- or part-time basis.

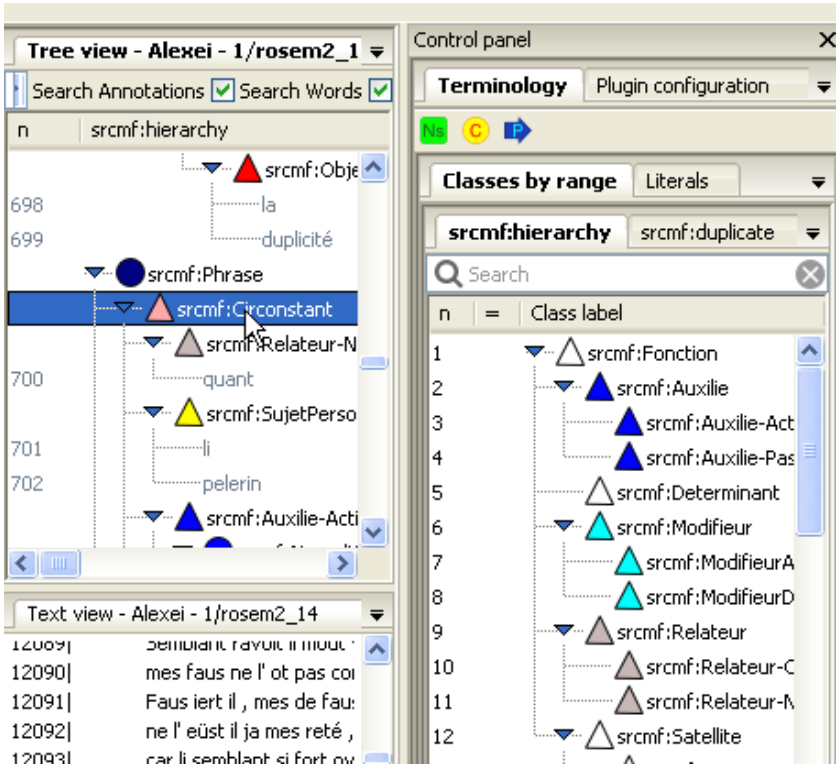

Fig. 1. NotaBene SRCMF annotation environment

All annotations are expressed using RDF formalism. They are stored separately from the XML-TEI source text files and are connected to the latter by using shared unique token identifiers. An SVN repository has been set up in order to manage the different versions of annotations and source texts.

3. Linguistic model

The syntactic model adopted by the project is dependency-based[1]. Syntactic relations are centered on inflected verb forms, which constitute the main governor nodes of clauses. Clauses with no finite verbs are either attached to the closest finite clause (in cases of coordination) or not analyzed. All the elements of the clause, including subjects, depend on the central verb. Morphosyntactic criteria are given a higher priority than semantic ones, although semantic criteria are used in some cases (for example to distinguish «essential» complements from optional «adjuncts»).

Two levels of annotation are implemented in the SRCMF project. The basic level is limited to the clause structure, the relations at word- and phrase-level are not analyzed except for identification of conjunctions and prepositions used to mark various kinds of relations (coordination, complement to verb, clause to clause). The boundaries of relative clauses are also marked, as their internal structure needs to be analyzed.

Deep level annotation marks all the syntactic relations including the internal structure of noun phrases.

Formulating exact definitions of annotated categories and instructions for dealing with various constructions occurring in the texts of the corpus has been a complicated task as the members of the project are influenced different schools of linguistic thought and also as methods such as acceptability judgments and fine semantic analysis are hardly applicable to a language that has no living native speakers. The comprehensive annotation manual is still under construction but will soon be available at the project website.

[1] *Polguère A., Mel'čuk I.* (ed.) Dependency in linguistic description. Amsterdam, Philadelphia. 2009.

4. Using the corpus

The SRCMF project does not intend to develop specific software for searching the corpus but relies on existing standard Treebank query tools.

SRCMF annotations can currently be exported to TigerXML format for use with TigerSearch software[1]. Web-based corpus query software is currently being tested thanks to integration of TigerSearch engine to the TXM text search and analysis platform[2].

Another export script converts SRCMF annotations to CoNLL defined by the Conference on Computational Natural Language Learning (e.g. CoNLL 2009 shared task) as the standard format for dependency parsers.

Due to copyright issues relating to some of the editions of Old French texts included in the corpus, it cannot currently be freely distributed to its users. However, all efforts are made to make access to the corpus as easy as possible.

5. Future developments

Future developments of the corpus include using the manual annotation as gold-standard training corpus for dependency parsers. Some tests have already been performed with *mate-tools*[3].

---

[1] **http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch**; *Lezius W.* TIGERSearch – Ein Suchwerkzeug für Baumbanken // Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002). Saarbrücken. 2002. **http://konvens2002.dfki.de**.

[2] **http://textometrie.ens-lyon.fr**; *Heiden S.* The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme // Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24). Sendai. 4–7 Novembre 2010. P. 389–398.

[3] *Björkelund A., Bohnet B., Hafdell L., Nugues P.* A high-performance syntactic and semantic dependency parser // Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010):

Both dependency parsing and manual annotation can be integrated into TXM platform which offers a convenient environment for qualitative and statistical corpus analysis.

Demonstrations Beijing. 2010. P. 33–36; *Bohnet B.* Top accuracy and fast dependency parsing is not a contradiction // Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing. 2010. P. 89–97.